IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

In re patent application of

Frederick J. Damerau and David E. Johnson

Serial No. 09/605,709

Filed June 27, 2000

For AUTOMATED SET UP OF WEB-BASED NATURAL LANGUAGE INTERFACE (As Amended)

Group Art Unit 2654

Examiner Abul K. Azad

Confirmation No. 3738

Commissioner for Patents
PO Box 1450
Alexandria, Virginia 22313-1450

APPELLANT'S SUPPLEMENTAL BRIEF UNDER 37 C.F.R. §1.193

This brief, which is filed herewith in triplicate, is responsive to a non-final rejection mailed March 29, 2004 reopening the prosecution of this case. The Applicant requests reinstatement of the appeal.

This brief contains additional statements and argument supplementing the original brief as indicated below. For convenience, the Summary of the Invention and the Appendix of Claims are repeated:

- ☐ I. REAL PARTY IN INTEREST
- ☐ II. RELATED APPEALS AND INTERFERENCES
- ☐ III. STATUS OF CLAIMS
- ☐ IV. STATUS OF AMENDMENTS

- ☑ V. SUMMARY OF INVENTION
- ☑ VI. ISSUES
- ☐ VII. GROUPING OF CLAIMS
- ☐ VIII. ARGUMENTS
  - ☐ ARGUMENT VIIIA. REJECTIONS UNDER 35 U.S.C. §112, FIRST PARAGRAPH
  - ☐ ARGUMENT VIIIB. REJECTIONS UNDER 35 U.S.C. §112, SECOND PARAGRAPH
  - ☑ ARGUMENT VIIIC. REJECTIONS UNDER 35 U.S.C. §102
  - ☐ ARGUMENT VIIID. REJECTIONS UNDER 35 U.S.C. §103
  - ☐ ARGUMENT VIIIE. REJECTION OTHER THAN 35 U.S.C. §§102, 103 AND 112
- ☑ IX. APPENDIX OF CLAIMS INVOLVED IN THE APPEAL
- ☐ X. OTHER MATERIALS THAT APPELLANT CONSIDERS NECESSARY OR DESIRABLE

V. SUMMARY OF INVENTION

The invention as defined in the claims on appeal is directed to a procedure that automates the process of setting up an instance of a conversational natural language interface for a Web site. By automating the process of setting up a new Web site, anyone can create a new interface. Subsequent manual tuning of the interface is possible and much easier to do than creating an interface from scratch. The invention solves the problem by bringing together a number of ideas and techniques, some of which have been used in natural language processing for other purposes. In order to set up an instance of a natural language conversational interface (NLCI), it is necessary to

1) define a hierarchy of topics into which individual documents or Web pages can be classified,
2) provide a keyword index for those documents for an associated search engine, and
3) for each node in the hierarchy, specify a mechanism for associating an input natural language (NL) query to the node. (In the preferred embodiment, this mechanism is a rule set and associated rule applier.)

To solve step (1), Applicants noted that the uniform resource locators (URLs) of the Web pages associated with a single site are often organized into a coherent hierarchy of topics. On reflection, this was not surprising, since good Web design encourages logical movement from page to page. Thus, a bank might have a Web page with the URL "www.bank.com/loans". It will have links to pages with URLs "www.bank.com/loans/auto" and "www.bank.com/loans/homemortgage", and so forth. (The URLs "www.bank.com/loans", "www.bank.com/loans/auto" and "www.bank.com/loans/homemortgage" are hypothetical for this example.) This is clearly a topic hierarchy of exactly the kind necessary for establishing the NLCI, in which "loans" is a high level node and "auto" and "homemortgage" are nodes

subordinate to it. If these are the lowest level in the hierarchy, the Web pages they point to are leaves.

To solve step (2), Applicants use methods from statistical natural language processing. From each document, Applicants generate a set of single words, bi-grams, etc., up to n-grams for some number *n*. However, these are not necessarily sequential n-grams. Applicants allow gaps between the words making up the n-gram. The term "sparse n-gram" is used at places in the patent application to emphasize the possibility that there might be gaps between the words in the n-gram. The concept of "sparse n-grams" as introduced by Applicants is unique to this patent application. The gaps between words are limited by establishing a distance *d* which is the maximum separation between the first and last words of the n-gram. This tactic is partial compensation for the variability allowed by natural language in expressing phrases. For example, one can say "input documents", or one might say "input text documents". The method described would generate an n-gram "input documents" from both of these. (In the preferred embodiment, words are reduced to stems, so the actual n-gram generated would be "input document".) The most frequent n-grams occurring in a document, up to some number *m*, are used as the keyword index for the document.

Figure 1 is a flow diagram of the automated set up procedure according to the invention. A program implementing a Web crawler is invoked in function block 11, beginning at the home page of the site for which a natural language interface is to be generated. The output of this module is a file of Web pages in HyperText Markup Language (HTML). In function block 12, the Uniform Resource Locators (URLs) of the Web pages are processed to induce a hierarchy of topics for the site and the HTML formatted pages are converted to the appropriate standard format. In a preferred implementation of the invention, the standard format is eXtensible Markup Language (XML). In function block 13, sparse n-grams are extracted from each page to serve as index terms for the page. The index terms are used to set up an answer

generator (search engine) for the page in function block 14. In function block 15, a set of sparse n-grams is generated for each of the topics found in function block 12 by grouping together all the documents having that topic. Those n-grams satisfying some criterion for significant association with the topic are saved. In a preferred implementation of the invention, the criterion used is the chi-square measure. The sparse n-grams are converted to rules in which each term of the n-gram is a term in the rule, and the topic is the rule consequent, in function block 16. Optionally, another statistical test can be made to associate a confidence measure with each rule. In the preferred implementation of the invention, the confidence measure is the percentage of time the underlying n-gram occurs in the topic. Once the preceding steps have been accomplished, all the necessary data is at hand to finish setting up the natural language interface in function block 17.

Figure 2 shows the components of the system and their inter-relationships. These include the Web crawler module 21 which begins at some designated home page(s) and systematically finds all the pages reachable from these initial pages, recursively. Using the URLs of these pages, module 22 finds the topic hierarchy of this site. Note that there might be more than one root (i.e., initial home page) resulting in more than one rooted tree (hierarchy). If there is more than one rooted tree, then the final hierarchy is just

top

$\swarrow \quad \searrow$

$root_1 \cdots root_n$

with new top node "Top_n. Module 23 uses the extracted pages along with the hierarchy to find key words and sparse phrases which can serve as index terms for the respective pages. Module 24 is an optional module for manual review and change of the decisions made by the automated system. Module 25 is a rules generating module which generates rules for each of the topics identified by module 22. Module 25 also uses the documents generated by the Web crawler module 21. The rules generated by

module 25 may optionally be edited manually, as indicated by the interface between modules 24 and 25. Module 26 is the interface builder system which uses the outputs of modules 23, 25 and, optionally, 24.

VI. ISSUES

The rejection mailed by the Examiner on 03/29/2004 raises the new issue whether claims 1 to 6 are anticipated by U.S. Patent No. 5,819,220 to Sarukkai et al. under the objective standards of 35 U.S.C. §102(b). This is the sole issue addressed in this Supplemental Appeal Brief.

ARGUMENT VIIIC. REJECTIONS UNDER 35 U.S.C. §102

The Examiner argues, upon reopening the prosecution, that the Sarukkai reference – which had been cited and argued in the prior §103 rejection – supports a §102 rejection. For the reasons which follow, some of which are repeated from the §103 section of the argument in the prior brief pertaining to Sarukkai, this argument cannot be sustained.

Sarukkai deals with a voice activated browser. In large part, Sarukkai deals with how to overcome problems with speech recognition algorithms when there are words that are "out of vocabulary". Instead of employing a rewriting style grammar, which is non-probabilistic and very rigid, Sarukkai employs n-grams. But n-grams also have the problem that they are statically trained on a given corpora and the Web will always have many words not in the training corpus, which means the speech recognition system.

To review the claimed invention, the basic set up is the following:

1. The system implicit in the invention, to which the automated set up methods pertain, requires a taxonomy of topics for a collection of documents, assumed to be associated with URLs, and a set of classification rules for each topic. The classification rules are used to classify user queries into topics as described in the now issued patent No. 6,567,805, cited as patent application Serial No. 09/570,788 in the cross-reference to related applications on page 1 of the specification.
2. The claimed invention specifies how to induce a taxonomy from a set of URLs and their associated documents and then a set of classification rules for the nodes in the taxonomy.
3. The method consists of (i) crawling a particular Web site, producing a set of Web pages (the documents to be associated with a taxonomy); (ii) using the structure of the URLs as the structure of the hierarchy; (iii) extracting from individual documents and from groups of documents, so-called sparse n-

grams, each of which is characteristic of a document or group of documents, where each group is associated with a node in the taxonomy; (iv) determining which phrases, whether sparse or not, are characteristic of the document or group of documents by some statistical technique for identifying salient collocations; and (v) converting the so-called sparse n-grams to classification rules for use in a classifier as described in patent No. 6,567,805 (cross-referenced as application Serial No. 09/570,788).

Note that the term "sparse n-gram", as defined and used in the disclosed and claimed invention, are sequences of tokens or words from the text where the tokens or words may or may not have other words between them. Perhaps the term "sparse n-gram" has confused the Examiner into thinking that the n-grams as used in the art of speech/voice recognition is relevant to the claimed invention. However, both the specification as filed and the foregoing explanation have made clear that the claimed invention is using the concept of n-grams in a different way than used in the art of speech/voice recognition. All that is meant is the more generic notion of a set (or sequence) of not necessarily adjacent tokens or words in the text. So for instance, in a document about mortgage loan applications, which has the phrase "mortgage loan application" as distinctive, one would presumably identify the phrase "mortgage loan" or even the noncontiguous phrase "mortgage application" as characteristic of the document.

As defined in the specification, "sparse n-gram" refers to a technique which allows for gaps between words making up the n-gram. The gaps are limited by establishing a distance *d* which is the maximum separation between the first and last words of the n-gram. The term "sparse" is taken from the reference cited in the specification at page 4, lines 2-5, which is a reference to the problem of "sparse data" (i.e. bi-grams which individually have such small probabilities that they don't show up in training data but have a significant probability of occurrence in the aggregate). While this connection has no intuitive relation to the definition of "sparse n-gram" in

the specification, the practical consequence is that one skilled in the art would have to look to the specification for the definition, which is proper since the applicant is entitled to be his own lexicographer. It is important to emphasize that the definition contained in the specification (page 3, lines 15-27) and used in the claims ("wherein the n-grams may be **sparse** or **non-sparse** n-grams", emphasis supplied) provides not only a definition unique to the present invention but also a very clear distinction, for those skilled in the art, between the use of "n-grams" for voice recognition (as in Sarukkai) and the use of "n-grams" in the present invention. Note also that there are two subcases of determining distinctive collocations (sparse n-grams): those distinctive of a single document and those distinctive of a group of documents. Many methods for doing this are well understood in the art and which is used is not material to the general idea of the disclosed and claimed invention.

Sarukkai simply does not deal with any of the topics addressed in the disclosed and claimed invention. The present invention and Sarukkai have in common use of the term "n-gram", but at a technical level these uses are quite distinct. For Sarukkai, "n-gram" means a sequence of tokens that are assigned probabilities within the context of a speech recognition system language model, which is irrelevant to the claimed invention. Many systems use common technologies, but even here the details of usage are very different. One cannot reasonably maintain that Sarukkai anticipates or teaches any features the claimed invention.

Specifically, Sarukkai does not mention using a taxonomy of topics let alone inducing a taxonomy. As the current invention is not about the specific use of the taxonomy or classification rules (this is covered in patent No. 6,567,805 cross-referenced as patent application Serial No. 09/570,788) and none of the cited references or patents mention this, it can be seen that they do not say anything relevant about this key part of the invention.

Nor does Sarukkai mention using so-called sparse n-grams in the manner used in the current invention, namely, in conjunction with documents and groups of documents associated with nodes or topics in an (induced) hierarchy to identify collocations or phrases that are characteristic of the associated document or group of documents.

Nor does Sarukkai mention converting sparse n-grams or collocations into classification rules, whose use is described in the context of a classification-based natural language interface for the Web in patent No. 6,567,805 (cross-referenced as application Serial No. 09/570,788).

It follows from this that Sarukkai does not deal in any way with the combination of these methods nor is such combination implicit in Sarukkai. It certainly cannot be reasonably maintained, when this is understood, that the claimed invention is anticipated by Sarukkai.

Sarukkai does teach the use of n-gram language models. However, the teachings of Sarukkai are not applicable to the claimed invention because they are not directed toward the set-up of a natural language interface. Sarukkai instead teaches methods for dynamically altering language models according to word sets in the documents searched. In other words, the language model is adjusted in response to documents found in a search. The n-grams used by Sarukkai are used for speech recognition, as known in the art. For example, Sarukkai teach smoothing or re-estimating "n-gram *language model scores*..." (col. 9, lines 20–21, emphasis added), thereby implying that the n-grams are used for speech recognition. N-grams are extremely well known in the art of speech recognition. By comparison, the n-grams employed in the present invention are created from documents to be searched, and the n-grams are stored as an index for searching. Hence, the n-grams in the present invention are used for very different purposes compared to the n-grams of Sarukkai

Sarukkai is concerned with augmenting or altering language / acoustic models, which is a specific part of statistical speech or voice recognition systems.

His invention is meant to improve speech or voice conversion of spoken language to text, in order to improve website navigation, browsing or ordinary text search. This is accomplished by altering the weights/parameter values of a language / acoustic model based on the occurrence of words, what is called a "web triggered word set" from "a selected subset of information in the document". Note that the "document" referred to by Sarukkai is whatever specific, single web page that the user of the speech (or voice) recognition-based search system is looking at. The main idea is that it is valuable to dynamically extract words from this page to bias the acoustic or language model in the speech recognition system.

One of the key notions in Sarukkai is that of a "web-triggered word set". This is not precisely defined but his discussion uses primarily if not exclusively words from links in web pages. The entire discussion is really based on navigation or browsing. It is possible that the Examiner has been misled by the frequent use of the word "search" in Sarukkai. For example, Sarukkai states that the "set of words constituting the link referent can constitute a web triggered word set, and it would make sense to *bias the speech recognition search* towards this set of words ..." (col 7, line 19-20) (emphasis added). It is essential to understand here that the word "search" in "the speech recognition search" does not refer to ordinary search as understood by someone using a search engine. "Search" as used in the quote and various other places in the Sarukkai invention in similar contexts refers to the highly specific and technical notion of searching through a language or acoustic model search space performed by a speech or voice recognition engine in the process of attempting to convert continuous spoken language input correctly to discrete textual output. That is, the Sarrukai invention is concerned with improving the accuracy of speech or voice recognition algorithms by using external information gathered in the course of an ordinary search of web pages (or other documents).

Sarukkai also states that "during the *search for the 'optimal' word sequence*, an evaluation function is used in all speech recognizers" (col 4, lines 58-59). Further,

" ... the web-triggered word set approach enables biasing the *search for the 'best' word sequence process* towards the set of words which is dynamically produced .... This word set biasing is achieved by incorporating the web-triggered word set as an additional piece of information in computing the speech recognition scores for the different word paths/sequences" (col. 5, line18 - 26). All of this refers to the notion of searching through the language or acoustic model internal to the speech recognition system and not to ordinary searching of web pages or other documents, which is a function from input text (the textual search terms or keywords) to textual keyword search indices of textual documents. Note also where Sarukkai states that the "speech recognizer then boosts the probability of these web triggered word sets, while determining the best word sequence. This word sequence is then matched with the stored word sequences corresponding to all the links possible from the currently viewed HTML document, in order to determine the next best link to take" (col 9, lines 36-41).

It could well be that the use of the term "search" – which has in the context of Sarukkai two utterly distinct senses as shown by example above – has confused the examiner into thinking the current invention and Sarukkai deal with related topics. However, the current invention (i) does not deal with speech or voice recognition systems or their acoustic / language models let alone biasing such models to improve speech recognition accuracy (although the Sarukkai methods could be used with the invention as an input device) and (ii) does not dynamically treat web pages as special in virtue of the fact that a user is currently looking at that particular page. In the current invention, all the processing of web pages that is done to set up a natural language interface is accomplished before the system is used. The current invention simply has no notion of a web page or document that a user is currently looking at. And Sarukkai has no notion of processing web pages in the sense done in the current invention. In the current invention the processing of web pages is crucially done in the context of a collection of documents or web pages, topic hierarchy and a mapping

of the collection to the topic hierarchy, as explained in the specification of the invention. In citing Sarukkai, the examiner is comparing apples to oranges.

"A claim is anticipated only if each and every element as set forth in the claim is found, either expressly or inherently described, in a single prior art reference." *Verdegaal Bros. v. Union Oil Co. of California*, 814 F.2d 628, 631, 2 USPQ2d 1051, 1053 (Fed. Cir. 1987). It may be helpful to reiterate, at greater length, that Sarukkai fails to describe the elements set forth in the claims of the present invention.

Claim element: "defining a hierarchy of topics ..."

The Examiner asserts that this claim element reads on Table 1, col. 7, lines 17-60. This is plainly in error. It is clear that the "triggered word set" defined by Sarukkai does not have a hierarchy. While it is well known that web pages are typically related in a hierarchical structure, the Sarukkai disclosure makes no use of this observation. Indeed, it is only the well known structure that is evident in Table 1. It could be said that a "hierarchy" of web pages is implicit in Table 1, but the claim element is not the hierarchy but rather "defining a hierarchy of topics." Only through improper hindsight can this claim element "read on" the Sarukkai disclosure.

There are two issues. First, the present invention can in many cases derive a useful notion of topic hierarchy from web sites. That observation is one part of the invention. The key insight or novelty is the use of that observation as explained in the specification. The examiner seems to be benefitting from hindsight: the current invention observes one can derive topic hierarchies from website structures and he retroactively sees a topic hierarchy in Table 1. Second, whether or not there is a topic hierarchy implicit in Table 1, Sarukkai makes absolutely no explicit use of that information. All that Sarukkai is doing is using the text information on the links that connect web pages to bias the acoustic or language model as explain in the Sarukkai patent. In his Table 1, the content under "Referent" clearly refer to the words on web page links or to common variations of those words, e.g., "tech" and "technology"; this is quite clear too from col 7, lines 28-35. Any word that occurs on a link, even "of"

is concerned relevant to Sarukkai. The reason for this is that the words showing up on a link in a web page are very likely to be uttered by someone using a speech recognition engine in the course of browsing a website. E.g., if you are on page X on a website and want to get to the previous web page, you are very likely to utter the word "back" or the phrase "go back" and hence you would like the acoustic or language model to be biased to expect those words in contrast to other words not occuring on a link or on the web page at all.

In contrast, in the current invention, extracting and establishing a topic hierarchy from a website is a essential aspect of the invention as classification rules are then derived that will map text queries to topics in the hierarchy. Sarukkai does not derive a topic hierarchy, does not derive – in the sense of the current invention – a set classification rules and hence *a fortiori* does not establish a set of classification rules per topic in a topic hierarchy.

Claim element: "generating a keyword index for those documents ..."

The Examiner asserts that this claim element reads on Sarukkai, col. 7, lines 17-60:

> "the information shown in the table was extracted automatically by a simple parsing JAVA program shown in Appendix 1. The set of words constituting the link referent can constitute a web triggered word set, and it would make sense to bias the speech recognition search towards this set of words since it is likely that the user will utter them."

The examiner's remarks seem inapposite. Sarukkai simply does not deal with indexing documents, where the index is to be used for a document search. Sarukkai only deals with extracting words from documents to bias an acoustic or language model of a speech or voice recognition system. The cited comment in Sarukkai could be interpreted to show that the examiner does not understand what indexing documents for (ordinary) search means or what Sarukkai is discussing, or both. The examiner is apparently being mislead by the fact that both inventions use the word "keyword" but with very different meanings.

Claim element: "for each topic in the hierarchy ..."

The Examiner asserts that this claim element reads on Sarukkai, col. 9, lines 17-24 and col. 10, lines 16-24. This is incorrect, as implicitly addressed above. As pointed out, Sarukkai does not use or even mention the notion of topic hierarchies. Sarukkai, in col. 9, lines 17-24, is concerned with biasing an acoustic or language model so that it would recognize with greater accuracy the words occurring on links on a web page that a user is currently looking at. The notion of n-gram in Sarukkai, as we have argued before, is not the same as the one used in the current invention. Sarukkai is concerned with n-grams as they are used in acoustic or language models and in particular Sarukkai is concerened with estimating probabilitistic weights for these n-grams. This is clear from the end of the quote provided by the examiner from Sarukkai, viz. " ... One method would be to appropriately smooth/re-estimate n-gram lanuage model scores ...." The same point applies to col 10, lines 16-24: cf. "the selective web triggered word set probability boosting just biases the speech recognition search ..." Again note that "search" in the previous quote means search through the acoustic or language model space to find the text words corresponding to the speech input and not to ordinary search of documents. None of this discussion in Sarukkai is relevant to the current invention.

Claim 2: "wherein the step of generating ...."

The Examiner asserts that this claim element reads on Sarukkai, at col 9, lines 19-22, and col. 10, lines 16-24 ("n-gram language model score using the HTML sources of the documents recently viewed"). This is incorrect. The examiners juxtaposition of the quote from the current invention and the citation from Sarukkai is simply absurd. What does generating an ordinary search keyword index have to do with "n-gram language model score..."? The answer is- there is no connection whatsoever.

Claim 3

The Examiner asserts that the claim limitation added in claim 3 reads on Sarukkai at col. 6, lines 36-39 ("modify the appropriate language Model and/or

acoustic model paramters dynamically in step 34, using the selected word-set list (step 32), to be used during the speech recognition search process"). This also evidences confusion. The review and possible modification mentioned in the current invention is manual and it is sensible to do this, i.e., a person could do this because the output are classification rules that people can understand. This is not the case with parameters of language and/or acoustic models, as in Sarukkai. It does not make sense to think one would or could manually review and modify a language Model and/or acoustic parameters. Such parameters are necessarily done with statistical estimation techniques.

Claim 4, the preamble

The Examiner asserts that the preamble of claim 4 reads on Sarukkai's teaching of "an automated method for setting up ..." in col 3, line 56 to col. 4, line 7. Once again the examiner makes the same basic mistake as there is no connection to the current invention, as explained above and in connection with claim 1. Sarukkai, as shown by the quote provided by the examiner, does not deal with setting up automatically an instance of a natural language interface in the sense of the current invention. As repeatedly pointed out to the examiner and amplified above, Sarukkai is only talking about a speech or voice recognition and altering parameters of the models in such a system.

Claim element: "automatically inducing a topic hierarchy ..."

The Examiner's assertion is answered by the discussion of this same claim element in claim 1, above.

Claim element: "creating rules from the n-grams ..."

The Examiner asserts that this claim element reads on Sarukkai, col. 7, lines 17-60, and col. 8, lines 54-67. This is incorrect for two reasons. First, n-grams as used in speech modeling are distinct from the n-grams discussed in the present invention, as discussed above. Second, Sarukkai does not create rules of any kind

from n-grams. Rather he is using words extracted from documents to smooth parameters in a language or acoustic model.

Claim 5

The Examiner asserts that the limitation added in claim 5 reads on the Sarukkai teaching "wherein the step of creating rules for a classification engine .." found at col 10, lines 10-15. This is incorrect. First, Sarukkai does not use rules, which he is criticizing in col 10, line 9. Cf. col 3, lines 5-7. Sarukkai's invention is meant to be an alternative to writing grammar rules. Cf. col 3 lines 47-54, viz., dynamically updating statistical models. Second, the grammar rules referred to/criticized in Sarukkai are distinct from the topic classification rules discussed in the present invention. The grammar rules for speech recognition are not topic classification rules at all. They are rules for recognizing grammatical phrases or patterns in language to constrain the output of a speech recognizer so that it is grammatical and likely. This is completely unrelated to the concerns of the current invention.

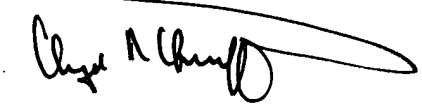
Claim 6

The response to the Examiner's assertion is covered by the above response to the Examiner's assertion regarding the limitation contained in claim 2.

Overall the examiner seems to be make identifications based on the use of the same word, out of context (n-gram, keyword, search, rule) and to arbitrarily juxtapose parts of the two unrelated inventions based on these superficial word identifications. As demonstrated above, this does not make a *prima facie* case for anticipation.

In summary, it is respectfully submitted that the claimed subject matter cannot properly be considered to be anticipated by the Sarukkai reference. As a matter of law, the Examiner's rejection on new §102 grounds should be reversed.

Respectfully submitted,



Clyde R Christofferson
Reg. No. 34,138

Whitham, Curtis & Christofferson, P.C.
11491 Sunset Hills Road, Suite 340
Reston, VA 20190
703-787-9400
703-787-7557 (fax)

Customer No. 30743

IX. APPENDIX OF CLAIMS INVOLVED IN THE APPEAL (37 C.F.R. §1.192(C)(9))

The text of the claims involved in the appeal are as follows:

1. An automated method for setting up a natural language interface in a Web site comprising the steps of:

defining a hierarchy of topics into which individual documents or Web pages can be classified;

generating a keyword index for those documents; and

for each topic in the hierarchy, associating a set of n-grams to a topic in the topic hierarchy, which set of n-grams is distinctive to that topic and wherein the n-grams maybe sparse or non-sparse n-grams.

2. The automated method for setting up a natural language interface in a Web site recited in claim 1, wherein the step of generating a keyword index comprises the step of extracting sparse n-grams of keywords for each group of pages in the topic hierarchy.

3. The automated method for setting up a natural language interface in a Web site recited in claim 1, further comprising the step of optionally reviewing and editing the keyword index.

4. An automated method for setting up a natural language interface in a Web site comprising the steps of:

automatically inducing a topic hierarchy by examining a structure of the Web site;

creating n-grams from pages in the Web site that are associated with a topic in the topic hierarchy wherein the n-grams may be sparse in-grams or non-sparse n-grams; and

creating rules from the n-grams, wherein each topic has associated rules that are used to decide if a new input document or query references the topic.

5. The automated method for setting up a natural language interface in a Web site recited in claim 4, wherein the step of creating rules is performed automatically and further comprising the optional step of manually editing the rules.

6. The automated method for setting up a natural language interface in a Web site recited in claim 1, further comprising the step of converting the set of n-grams to classification rules.